PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-305856

(43)Date of publication of application : 02.11.2000

(51)Int.Cl. G06F 12/16
G06F 3/06
G06F 11/20
G06F 12/00
G06F 13/00

(21)Application number : 11-117670

(22)Date of filing : 26.04.1999

(71)Applicant : HITACHI LTD

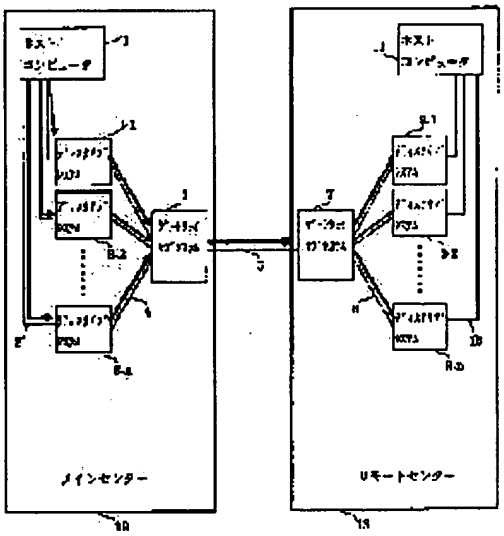
(72)Inventor : TABUCHI HIDEO
NOZAWA MASASHI
SHIMADA AKINOBU

(54) DISK SUBSYSTEMS AND INTEGRATION SYSTEM FOR THEM

(57)Abstract:

PROBLEM TO BE SOLVED: To guarantee the sequence of update and the consistency of data by doubling data between disk subsystems on a main-center side and a remote- center side through gateway subsystems.

SOLUTION: Data which are written from a host computer 1 are doubled between disk subsystems 3-1, 3-2...3-n and a gateway subsystem 5 and held macroscopically in the same state. The gateway subsystem 5 adds information for holding the sequence of update. Further, the data are doubled between the gateway subsystem 5 and a gateway subsystem 7 by asynchronous remote copying while the sequence of update is guaranteed. The disk subsystems 9-1, 9-2...9-n have the data updated in synchronism with the update of the gateway subsystem 7. Those are all actualized only by the function of the disk subsystems and no new software need not be introduced.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] It is the disk subsystem which it is the disk subsystem connected to the 1st external storage and the 2nd external storage, and data transfer between said 1st external storage is a synchronous type, and data transfer between said 2nd external storage is an asynchronous type, and is performed, respectively.
[Claim 2] It is the integration system which it is the integration system which has a disk subsystem connected to high order equipment, the 1st external storage connected to said high order equipment, and said the 1st external storage and 2nd external storage, and data transfer between said 1st external storage of said disk subsystem is a synchronous type, and data transfer between said 2nd external storage is an asynchronous type, and is characterized by carrying out, respectively.
[Claim 3] Said 2nd external storage is a disk subsystem according to claim 1 characterized by said 2nd external storage existing in a remote place, or an integration system according to claim 2 characterized by existing in a remote place.
[Claim 4] It is the integration system according to claim 2 connected through a communication line between a disk subsystem according to claim 1 with which it connects through a communication line between said 2nd external storage or said disk subsystem, and said 2nd external storage.
[Claim 5] An integration system according to claim 2 a disk subsystem according to claim 1 said whose the 1st and said 2nd external storage are a disk subsystem, respectively or said 1st [the], and said whose 2nd external storage are disk subsystems, respectively.
[Claim 6] High order equipment The 1st external storage connected to said high order equipment The Maine pin center,large which has a disk subsystem connected to said the 1st external storage and 2nd external storage Said 2nd external storage connected to the 3rd external storage and said disk subsystem It is the integration system equipped with the above, and is characterized by updating sequence of data from said high order equipment turning into updating sequence of data over said 3rd external storage in a remote pin center,large.
[Claim 7] It is the integration system which data transfer between said 1st external storage and said disk subsystems is a synchronous type, and data transfer between said 2nd external storage and said disk subsystems is an asynchronous type, and is performed in an integration system of said 6th publication, respectively.
[Claim 8] An integration system according to claim 2 with which information about sequence is added to data on the occasion of said asynchronous data transfer on the occasion of a disk subsystem according to claim 1 with which information about sequence is added to data, or said asynchronous data transfer.
[Claim 9] A disk subsystem according to claim 8 whose information about sequence added to said data is a serial number, or an integration system according to claim 8 whose information about sequence added to said data is a serial number.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which invention belongs] This invention connects mutually two or more external storage groups (disk subsystem) which exist in a remote place especially, and two or more of other external storage groups about the external storage which stores the data of a computer system, and these integration systems, and relates to the remote copy technology which doubles data between the external storage (disk subsystem) which exists in a remote place, without going via a high order equipment slack host computer. Here, a disk subsystem shall mean the storage which contains the control section which delivers and receives information to high order equipment, and the disk unit which performs informational storing.
[0002]
[Description of the Prior Art] The enternal memory system which doubles and holds data between the disk subsystems currently installed in the Maine pin center,large and the remote pin center,large, respectively and which adopted the so-called remote copy function is already put in practical use partly.
[0003] With this conventional technology, in order for a host computer to intervene and to attain the function of a remote copy, various technical problems occurred.
[0004] A "synchronous type and asynchronous type" remote copy function is divided roughly into two kinds, a synchronous type and an asynchronous type.
[0005] A synchronous type means the procedure which reports completion of an update process to the host computer of the Maine pin center,large, after directed updating (writing) is completed to the disk subsystem in the remote pin center,large which is the object of the remote copy function, when a disk subsystem has updating (writing) directions of data from the host computer in the Maine pin center,large (high order equipment) and the referent is also an object of a remote copy function. According to the geographical distance of the Maine pin center,large and a remote pin center,large, it is influenced of the capacity of the data transmission line which intervenes in the meantime, and time lags (transmission time etc.) occur. When the transmission time was taken into consideration, even if the synchronous type was called remote place, dozens of km was a limit actually.
[0006] In the synchronous type, the content of the data of the disk subsystem of the Maine pin center,large and a remote pin center,large sees macroscopically, and is always in agreement. For this reason, since the condition of a just before [disaster] is saved thoroughly at the disk subsystem by the side of a remote pin center,large and it is even if it is the case where the Maine pin center,large loses a function according to disaster etc., there is an effect which can resume processing promptly by the remote pin center,large side. In addition, by seeing macroscopically, coincidence always means that data is surely in the same condition at the event of the completion of a data update process, although it is the unit (microsecondec, msec) of the processing time of a magnetic disk drive or an electronic circuitry and the condition of not being in agreement is possible during operation in a synchronous function. This is because an update process of the Maine pin center,large cannot be completed, unless reflection of the updating data to a remote pin center,large is completed. For this reason, especially, the distance of the

Maine pin center,large and a remote pin center,large is separated, and when the data transmission line is congested, the access engine performance of the disk subsystem by the side of the Maine pin center,large deteriorates substantially.

[0007] On the other hand, as soon as an update process of the disk subsystem in the Maine pin center,large finishes even if the referent is an object of a remote copy function when a disk subsystem has updating (writing) directions of data from the host computer in the Maine pin center,large, an asynchronous type reports completion of an update process to a host computer, and means the procedure which performs processing [in / in the renewal of the data in the disk subsystem of a remote pin center,large (reflection) / the Maine pin center,large] to asynchronous. For this reason, since renewal of data is completed by the processing time needed inside the Maine pin center,large, the transmission time resulting from storing of the data to a remote pin center,large etc. does not start.

[0008] The content of an asynchronous type of the disk subsystem of a remote pin center,large does not always correspond to it by the side of the Maine pin center,large. For this reason, when the Maine pin center,large loses a function according to disaster etc., the data which reflection of data has not completed to a remote pin center,large side will disappear. However, the access engine performance of the disk subsystem by the side of the Maine pin center,large can be made into the case where a remote copy function is not carried out, and equivalent level.

[0009] If backup of the data in the case of natural disasters, such as an earthquake, is taken into consideration, it is necessary to separate the Maine pin center,large and a remote pin center,large 100km - about 100km of numbers. Moreover, although it is also possible to, use the high-speed communication line of 300 Mbit/sec classes from 100 Mbit/sec for example, for a remote copy function, the customer of a disk subsystem is made to pay the connection fees of a large sum, and it is not economical.

[0010] "Sequentiality maintenance" When it is going to back up the Maine pin center,large which has two or more disk subsystems other than the technical problem of the transmission time of data in the remote pin center,large, the technical problem that each disk subsystem must correspond to 1 to 1 (sequentiality maintenance) occurs. It is unavoidable that reflection of the updating data to a remote pin center,large is late for the generating event of a actual update process in the Maine pin center,large, and is processed by the asynchronous remote copy. However, the sequence of updating must be in agreement with the Maine pin center,large.

[0011] Generally, the data base etc. consists of a main part of a data base, various log information, and control information, and each has relevance. In the case of renewal of data, in addition to the main part of a data base, these log information and control information are also updated and the consistency of a system is maintained. Therefore, when the sequence of updating collapses, the consistency of such information relevant to updating sequence also collapses, and when the worst, it may lead to destruction of the whole data base.

[0012] In realizing a remote copy asynchronous in the general environment where two or more disk subsystems exist in a "host computer intervenes" Maine pin center,large and a remote pin center,large, when a host computer directs renewal of data to a disk subsystem, it is common that add the information about updating sequence, such as a time stump, to data, and updating data reflection processing of the disk subsystem of sub** is performed based on such information. For example, like the XRC (Extended Remote Copy) function of IBM, the host computer intervened and the remote copy function is realized.

[0013] Concrete disclosure of a XRC function is made by JP,6-290125,A (U.S. Pat. No. 5446871) at details. In a XRC function, cooperation of the operating system of the host computer by the side of the Maine pin center,large, a disk subsystem, and the data mover software of the host computer by the side of a remote pin center,large and a disk subsystem has realized issuance of updating sequence information, sending, and updating data reflection processing based on this.

[0014]

[Problem(s) to be Solved by the Invention] An asynchronous remote copy function is realizable with the conventional technology (XRC function), guaranteeing the updating sequentiality between the Maine pin center,large and a remote pin center,large. However, with the conventional technology, the structure for XRC functional implementation is required for the both sides of high order software and a disk

subsystem, and both have to cooperate. Since the new software of dedication needs to be introduced, the activity of reexamination of the system design accompanying installation of software, setting out, inspection, and the increment in a CPU load etc. generates a user. For this reason, the predetermined period was required for installation of the conventional function, and there was an introductory obstruction that costs occurred.

[0015] Moreover, when the capacity of a communication line with a remote pin center,large was not enough and the asynchronous remote copy function was performed, the technical problem that non-reflected updating data increased to a remote pin center,large occurred.

[0016] The object of this invention does not need installation of new software, but is only the function of a disk subsystem, can guarantee the sequentiality of updating, and the consistency of data, and is realizing an asynchronous remote copy function with little degradation of the Maine pin center,large with easy and installation.

[0017] Another object of this invention is applying an asynchronous remote copy function to the disk subsystem which can perform mass data storage, and it is to realize a remote copy function, without making the customer of a disk subsystem pay the connection fees of a large sum.

[0018]

[Means for Solving the Problem] One disk subsystem (a henceforth, gateway subsystem) which serves as the gateway, respectively is arranged in the Maine pin center,large which exists mutually in a remote place, and each remote pin center,large, and a gateway subsystem is connected to them on the data transmission line. And all disk subsystems that need to carry out a remote copy of both pin center,larges are connected to each gateway subsystem in a pin center,large.

[0019] Between volume set as the object of a remote copy of a disk subsystem of the Maine pin center,large, and volume of arbitration of a gateway subsystem in the Maine pin center,large, it connects by the synchronous remote function and data is doubled. Thereby, in volume of a disk subsystem for [in the Maine pin center,large] a remote copy, and volume of a gateway subsystem in the Maine pin center,large, if delay of the processing time of a system etc. can be disregarded, the same data will be held.

[0020] Between each volume of a gateway subsystem of the Maine pin center,large and a remote pin center,large, data is doubled by the asynchronous remote copy. However, according to sequence that volume in a self subsystem was updated, as for a gateway subsystem of the Maine pin center,large, updating data is reflected in volume in a self subsystem according to sequence from which updating data was sent to a gateway subsystem of a remote pin center,large, and a gateway subsystem of a remote pin center,large received it.

[0021] Data is doubled by the synchronous remote copy between each volume of a gateway subsystem in a remote pin center,large, and each disk subsystem. Thereby, in volume of a gateway subsystem by the side of a remote pin center,large, and volume of a disk subsystem for a remote copy, it sees macroscopically and the same data is always held.

[0022] In addition, data of volume for a remote copy is stored in buffer memory in a gateway subsystem of self in a gateway subsystem. Therefore, area of a subsystem for data storage is not usually necessarily needed other than buffer memory. However, if there is area of an available subsystem, area of a subsystem which is needed according to capacity of the transmission line in the case of data transmission and reception which went via the transmission line can be used.

[0023] It is realizable with the above configuration, maintaining the sequentiality of updating of doubleness of data of two or more disk subsystems of the Maine pin center,large, and two or more disk subsystems of a remote pin center,large only by function of a disk subsystem. Reflection of updating data to a remote pin center,large can be carried out to asynchronous with an update process of each disk subsystem of the Maine pin center,large. Thereby, easy disaster backup system of installation can be offered with high performance. Moreover, according to channel capacity of the transmission line, suitably, area of a subsystem can be used and there is an effect which can mitigate a customer's connection-fees burden.

[0024]

[Embodiment of the Invention] Hereafter, an example at the time of applying this invention to a general-purpose computer system is explained, referring to a drawing.
[0025] In two or more data centers which equipped drawing 1 with the general-purpose computer system, in order to double data between two pin center,larges of arbitration, the example of a configuration when applying this invention is shown.
[0026] One set or two or more sets of one set or two or more sets of the disk subsystems by the side of the Maine pin center,large, and the disk subsystems by the side of a remote pin center,large were connected through the Gateway subsystem, without minding a host computer, and the remote copy system which doubles data among both pin center,larges is realized.
[0027] the Maine pin center,large 12 of drawing 1 -- setting -- a central processing unit (host computer) 1 -- the interface cable 2 -- minding -- a disk subsystem 3-1, 3-2, and ..... it connects with 3-n. a disk subsystem 3-1, 3-2, and ..... 3-n stores the data referred to or updated from a host computer 1. The gateway subsystem 5 is connected with 3-n from a disk subsystem 3-1 through an interface cable 4.
[0028] The gateway subsystem 5 will write the data concerned also in the buffer memory in a self subsystem synchronizing with this, if a host computer publishes the write request of data in disk subsystem 3-1 grade. Furthermore, with data having been written in the buffer memory in a self subsystem, write-in directions of data are performed to asynchronous to the gateway subsystem 7 which exists in a remote place. The gateway subsystem 5 surely consists of disk subsystems 3-1 irrespective of the number of 3-n at one set.
[0029] The gateway subsystem 7 is installed in the remote pin center,large 13, and is connected with the gateway subsystem 5 of the Maine pin center,large 12 through the interface cable 6. In addition, the interface cable 6 can also be connected with a general communication line. This example describes as an interface cable 6 also including this point. The gateway subsystem 7 is stored in the buffer memory in a self subsystem at order with write-in directions of the data received from the gateway subsystem 5. The gateway subsystem 7 surely consists of one set.
[0030] a disk subsystem 9-1, 9-2, and ..... 9-n is connected with the gateway subsystem 7 through the interface cable 8. Disk subsystem 9-1 grade writes in the data concerned also in a self subsystem synchronizing with this, when the write request of data is in the gateway subsystem 7 from the Maine pin center,large 12.
[0031] That is, from a host computer 1, when there are write-in directions of data to 3-n from one set or two or more sets of disk subsystems 3-1, the same data also as 9-n is stored from one set or two or more sets of the disk subsystems 9-1 in the remote pin center,large 13. The arrow head of drawing 1 shows the data flow which had write-in directions from the host computer 1.
[0032] A host computer 11 is a central processing unit which is connected with 9-n by the interface cable 10 from a disk subsystem 9-1 in the remote pin center,large 13, and performs reference and updating to disk subsystem 9-1 grade. A host computer 11 can be processed by becoming an alternative of a host computer 1, when it becomes impossible for the host computer 1 of the Maine pin center,large 12 to achieve an original function by disaster, failure, etc. In addition, the data stored in the disk subsystem 9-1 grade can be used, and processing which is different in the host computer 1 of the Maine pin center,large 12 can be performed separately [a host computer 1] independently. However, when a host computer 11 does not process to disk subsystem 9-1 grade, the host computer 11 is unnecessary.
[0033] As a gestalt of operation of this invention, the doubleness method of data and the outline of employment are explained using drawing 2 and drawing 3 .
[0034] An employment person chooses in advance the volume in which the data set as the object of doubleness was stored, a data set, and a disk subsystem. And the employment person sets up beforehand the relation between object volume, an object data set and a disk subsystem, and the volume which stores the duplicate of selected data, a data set and a disk subsystem from the host computer to the disk subsystem.
[0035] The above-mentioned selection and setting out are faced, and in the case of the disk subsystem which can connect or equip the console and service processor of dedication, a host computer is not used, but it can set to it through the console and service processor. The flow of drawing 2 shows the case

where selection and setting out are performed from a host.
[0036] The method of specifying the concrete address meaning above-mentioned volume and an above-mentioned disk subsystem as the setting-out method, and the method of choosing from the range of the arbitration of the address with the control program in a disk subsystem can also be taken. As initial setting, the example which performs pass setting out and pair setting out is shown (drawing 2 , 201).
[0037] from a host computer 1 (drawing 1), the write request (henceforth, light command) of data publishes to a disk subsystem 3-1, 3-2, ..., 3-n (211) -- having (drawing 2 , 202) -- A disk subsystem 3-1, 3-2, ..., 3-n publish the light command of the data to the gateway subsystem 5 (212), performing data storage processing into a self disk subsystem based on a light command (203). Here, a light command is a command which transmits the directions for writing in data, and the write-in data itself.
[0038] The gateway subsystem's 5 receipt of a light command performs processing to a light command (204). If the data storage processing to the light command to the buffer memory in the gateway subsystem of self is completed, the gateway subsystem 5 will report completion of processing to a disk subsystem 3-1, 3-2, ..., 3-n (211). In connection with this, the light command number is given to the order which processing completed for every light command (205), and the light command with which the light command number was given is published to the gateway subsystem 7 (213) by the opportunity determined based on the throughput of a self subsystem to a light command numerical order (206).
[0039] On condition that the processing to the light command, i.e., data storage processing into a self subsystem, is completed to the disk subsystem 3-1, 3-2, ..., light command in which 3-n (211) was published from the host computer 1 and completion of write-in processing is reported from the gateway subsystem 5 (212) (221), the completion report (222) of the processing to a light command is performed to a host computer 1.
[0040] A check of that the gateway subsystem 7 (213) has received the light command to the given numerical order by the light command number given to the light command published from the gateway subsystem 5 (212) performs the processing to a light command, i.e., the data storage processing to the buffer memory in a self subsystem, (301). In connection with this, the light command of the data is published to a disk subsystem 9 (311) (302). A disk subsystem's 9 (311) receipt of the light command published from the gateway subsystem 7 performs the processing to a light command, i.e., data storage processing into a self subsystem, (303).
[0041] Completion of a disk subsystem 9-1, 9-2, ..., processing of as opposed to a light command in 9-n (311), i.e., the data storage processing to the buffer memory in a self subsystem, performs the completion report (321) of processing to the gateway subsystem 7. Data storage processing into a self subsystem completes the gateway subsystem 7 (213), and on condition that completion of write-in processing is reported [a disk subsystem 9-1, 9-2, ...,] from 9-n, the completion report (322) of processing to a light command is performed to the gateway subsystem 5.
[0042] It doubles a disk subsystem 3-1, 3-2, ..., between 3-n and the gateway subsystem 5, and the data written in from the host computer 1 is seen macroscopically, and is always maintained at the same condition by this invention. In this case, the information (serial number) for holding the sequence of updating in the Gateway subsystem 5 is added.
[0043] Moreover, doubleness of data is performed by the asynchronous remote copy between the gateway subsystem 5 and the gateway subsystem 7, guaranteeing the sequence of updating. As for a disk subsystem 9-1, 9-2, ..., 9-n, data is updated synchronizing with renewal of the gateway subsystem 7. Including the disk subsystem which has a Gateway function, it realizes only by the function of a disk subsystem and these all do not serve as a burden to the throughput of a host computer.
[0044] The buffer field in each Gateway subsystem is used for drawing 4 , and actuation when the channel capacity of the transmission line is not enough is explained to it. It is shown that the same sign is already explanation ending. In this system, the buffer field which stores data with a write request temporarily is prepared in each Gateway subsystem. It is for preventing that the buffer memory in the usual transmission line overflows. The data of the subsystem stored in this buffer field is sent to a remote pin center,large from the Maine pin center,large through the transmission line, and is inputted into a Gateway subsystem through the buffer field by the side of a remote pin center,large there. If it

carries out like this, although whenever [doubleness time coincidence] will decrease, ** can also realize an asynchronous remote copy function without a mass communication line.
[0045]
[Effect of the Invention] Installation of new software is not needed, but only by the function of a disk subsystem, the sequentiality of updating and the consistency of data can be guaranteed, installation is easy and an asynchronous remote copy system without lowering of the processability ability of the Maine pin center,large can be realized.
[0046] Moreover, according to the channel capacity of the transmission line, suitably, the area of a subsystem can be used and there is an effect which can mitigate a customer's connection-fees burden.

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－305856
（Ｐ2000－305856Ａ）
(43)公開日　平成12年11月2日(2000.11.2)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
Ｇ０６Ｆ	12/16	３１０	Ｇ０６Ｆ 12/16	３１０Ｍ	５Ｂ０１８
	3/06	３０４	3/06	３０４Ｆ	５Ｂ０３４
	11/20	３１０	11/20	３１０Ｃ	５Ｂ０６５
	12/00	５３１	12/00	５３１Ｄ	５Ｂ０８２
	13/00	３０１	13/00	３０１Ｒ	５Ｂ０８３

審査請求　未請求　請求項の数９　ＯＬ　（全 10 頁）

(21)出願番号　特願平11－117670

(22)出願日　平成11年４月26日(1999.4.26)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　田渕　英夫
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内

(72)発明者　野沢　正史
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内

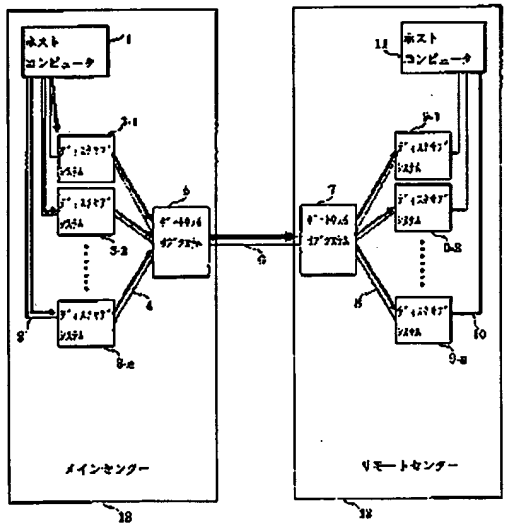
(74)代理人　100075096
弁理士　作田　康夫

最終頁に続く

(54)【発明の名称】　ディスクサブシステム及びこれらの統合システム

(57)【要約】　（修正有）
【課題】ホストコンピュータに負担を掛けることなく、複数のディスクサブシステムにわたってデータ更新の順序性／データの整合性を保証でき、導入が容易かつホストコンピュータの性能低下が無い、非同期型のリモートコピー機能を有するディスクサブシステムを提供する。
【解決手段】各センターのディスクサブシステムのリモートコピーの対象となるボリュームとゲートウェイサブシステムの任意のボリュームの間は同期型リモートコピーでデータの二重化が行なわれ、メインセンターのゲートウェイサブシステムは自サブシステム内のボリュームが更新された順番に従い更新データをリモートセンターのゲートウェイサブシステムに送付しリモートセンターのゲートウェイサブシステムは受け取った順番に従い更新データを自サブシステム内のボリュームに反映する非同期型のリモートコピーでデータの二重化が行なわれるリモートコピーシステム。

図1



1

【特許請求の範囲】

【請求項１】第１の外部記憶装置と．第２の外部記憶装置とに接続されるディスクサブシステムであって．

前記第１の外部記憶装置との間のデータ転送は同期型で．

前記第２の外部記憶装置との間のデータ転送は非同期型で．それぞれ行うディスクサブシステム。

【請求項２】上位装置と．

前記上位装置に接続された第１の外部記憶装置と．

前記第１の外部記憶装置及び第２の外部記憶装置に接続されたディスクサブシステムを有する統合システムであって．

前記ディスクサブシステムは、前記第１の外部記憶装置との間のデータ転送は同期型で、前記第２の外部記憶装置との間のデータ転送は非同期型で．それぞれ行うことを特徴とする統合システム。

【請求項３】前記第２の外部記憶装置は遠隔地に存在することを特徴とする請求項１記載のディスクサブシステム．又は、

前記第２の外部記憶装置は遠隔地に存在することを特徴とする請求項２記載の統合システム。

【請求項４】前記第２の外部記憶装置との間は．通信回線を介して接続されている請求項１記載のディスクサブシステム、又は．

前記ディスクサブシステムと前記第２の外部記憶装置との間は、通信回線を介して接続されている請求項２記載の統合システム。

【請求項５】前記第１及び前記第２の外部記憶装置が、それぞれ、ディスクサブシステムである請求項１記載のディスクサブシステム、又は、

前記第１及び前記第２の外部記憶装置が、それぞれ、ディスクサブシステムである請求項２記載の統合システム。

【請求項６】上位装置と．前記上位装置に接続された第１の外部記憶装置と、

前記第１の外部記憶装置及び第２の外部記憶装置に接続されたディスクサブシステムを有するメインセンターと．

第３の外部記憶装置及び前記ディスクサブシステムに接続された前記第２の外部記憶装置を有するリモートセンターからなる統合システムであって．

前記上位装置からのデータの更新順序が、リモートセンターにおける前記第３の外部記憶装置に対するデータの更新順序となることを特徴とする統合システム。

【請求項７】前記第６記載の統合システムにおいて、

前記第１の外部記憶装置と前記ディスクサブシステムとの間のデータ転送は同期型で、

前記第２の外部記憶装置と前記ディスクサブシステムとの間のデータ転送は非同期型で、それぞれ行う統合システム。

2

【請求項８】前記非同期型のデータ転送に際し．データに順序に関する情報が付加されている請求項１記載のディスクサブシステム、又は．

前記非同期型のデータ転送に際し、データに順序に関する情報が付加されている請求項２記載の統合システム。

【請求項９】前記データに付加された順序に関する情報がシリアル番号である請求項８記載のディスクサブシステム．又は、

前記データに付加された順序に関する情報がシリアル番号である請求項８記載の統合システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明はコンピュータシステムのデータを格納する外部記憶装置及びこれらの統合システムに関し．特に、遠隔地に存在する複数の外部記憶装置群（ディスクサブシステム）と．他の複数の外部記憶装置群とを相互に接続し．上位装置たるホストコンピュータを経由せずに、遠隔地に存在する外部記憶装置（ディスクサブシステム）との間で．データを二重化するリモートコピー技術に関する。ここで、ディスクサブシステムとは．上位装置に対し情報の授受を行う制御部と．情報の格納を行うディスク装置を内蔵する記憶装置をいうものとする。

【０００２】

【従来の技術】メインセンターとリモートセンターにそれぞれ設置されているディスクサブシステムの間で、データを二重化して保有する．いわゆる．リモートコピー機能を採用した外部記憶システムが．既にいくつか実用化されている。

【０００３】かかる従来技術では、ホストコンピュータが介在してリモートコピーの機能を達成するため．種々の課題があった。

【０００４】「同期型と非同期型について」リモートコピー機能は、同期型と非同期型の２種類に大別される。

【０００５】同期型とはメインセンター内のホストコンピュータ（上位装置）からディスクサブシステムに、データの更新(書込み)指示が有った場合．その指示対象がリモートコピー機能の対象でもあるときは、そのリモートコピー機能の対象であるリモートセンターにおけるディスクサブシステムに対して、指示された更新（書込み）が終了してから、メインセンターのホストコンピュータに更新処理の完了を報告する処理手順をいう。メインセンターとリモートセンターとの地理的距離に応じて．この間に介在するデータ伝送線路の能力の影響を受け．時間遅れ（伝送時間等）が発生する。同期型は、伝送時間を考慮すると遠隔地といっても現実的には数十Ｋｍが限界であった。

【０００６】同期型では．メインセンターとリモートセンターのディスクサブシステムのデータの内容が巨視的にみて常に一致している。このため．メインセンターが

災害等により機能を失った場合であっても、リモートセンター側のディスクサブシステムに災害直前までの状態が完全に保存されいるので、リモートセンター側で迅速に処理を再開できる効果がある。尚、巨視的にみて常に一致とは、同期型の機能を実施中には、磁気ディスク装置や電子回路の処理時間の単位（μsec、msec）で、一致していない状態が有り得るが、データ更新処理完了の時点ではデータは必ず同一の状態になっていることを意味している。これは、リモートセンターへの更新データの反映が終了しない限り、メインセンターの更新処理を完了できないためである。このため、特に、メインセンターとリモートセンターの距離が離れており、データ伝送線路が複雑している場合には、メインセンター側のディスクサブシステムのアクセス性能が大幅に劣化する。

【０００７】これに対し非同期型とは、メインセンター内のホストコンピュータからディスクサブシステムに、データの更新(書込み)指示が有った場合、その指示対象がリモートコピー機能の対象であっても、メインセンター内のディスクサブシステムの更新処理が終わり次第、ホストコンピュータに対し更新処理の完了を報告し、リモートセンターのディスクサブシステムにおけるデータの更新（反映）はメインセンターにおける処理とは非同期に実行する処理手順をいう。このためメインセンター内部で必要とされる処理時間でデータ更新が終了するので、リモートセンターへのデータの格納に起因する伝送時間等はかからない。

【０００８】非同期型は、リモートセンターのディスクサブシステムの内容が、メインセンター側のそれに対し、常に一致しているわけではない。このため、メインセンターが災害等により機能を失った場合は、リモートセンター側にデータの反映が完了していないデータが消失することとなる。しかし、メインセンター側のディスクサブシステムのアクセス性能を、リモートコピー機能を実施しない場合と同等レベルとすることができる。

【０００９】地震等の天災の際のデータのバックアップを考慮すれば、メインセンターとリモートセンターは１００ｋｍ～数１００ｋｍ程度、分離する必要がある。また、例えば、１００Ｍｂｉｔ／ｓｅｃから３００Ｍｂｉｔ／ｓｅｃクラスの高速通信回線をリモートコピー機能のために使用することも可能ではあるが、ディスクサブシステムの顧客に高額の回線使用料を負担させることとなり、経済的でない。

【００１０】「順序性保全について」データの伝送時間の課題の他に、複数のディスクサブシステムを有するメインセンターのバックアップをリモートセンターで行おうとするとき、各々のディスクサブシステムが１対１に対応しなければならない（順序性保全）という課題がある。非同期型リモートコピーでは、リモートセンターへの更新データの反映が、メインセンターでの実際の更新処理の発生時点より遅れて処理されることはやむを得ない。しかし更新の順序はメインセンターと一致していなければならない。

【００１１】一般にデータベース等はデータベース本体と各種ログ情報、制御情報から構成されており、それぞれが関連性を持っている。データ更新の際はデータベース本体に加え、これらログ情報、制御情報をも更新し、システムの整合性が保たれている。したがって更新の順序が崩れた場合、更新順序に関連するこれらの情報の整合性も崩れ、最悪の場合には、データベース全体の破壊につながる可能性がある。

【００１２】「ホストコンピューターが介在することについて」メインセンター及びリモートセンターに複数のディスクサブシステムが存在する一般的な環境で非同期型のリモートコピーを実現する場合には、ホストコンピュータがディスクサブシステムへデータの更新を指示する場合、タイムスタンプなどの更新順序に関する情報をデータに付加し、これらの情報に基づいて副側のディスクサブシステムの更新データ反映処理が実行されるのが一般的である。例えば、ＩＢＭ社のＸＲＣ（Ｅｘｔｅｎｄｅｄ Ｒｅｍｏｔｅ Ｃｏｐｙ）機能のように、ホストコンピュータが介在してリモートコピー機能を実現している。

【００１３】ＸＲＣ機能の具体的開示は特開平６－２９０１２５（米国特許第５４４６８７１号）に詳細になされている。ＸＲＣ機能においては、メインセンター側のホストコンピュータのオペレーティングシステムとディスクサブシステム、リモートセンター側のホストコンピュータのデータムーバーソフトウェアとディスクサブシステムの連携により、更新順序情報の発行、送付、これに基づく更新データ反映処理を実現している。

【００１４】

【発明が解決しようとする課題】従来技術（ＸＲＣ機能）により、メインセンター、リモートセンター間の更新順序性を保証しながら非同期型のリモートコピー機能が実現できる。しかし従来技術では、上位ソフトウェアとディスクサブシステムの双方にＸＲＣ機能実現の為の仕組が必要であり、且つ、両者が連携しなければならない。専用の新規ソフトウェアの導入が必要なため、ユーザは、ソフトウェアの導入、設定、検査、ＣＰＵ負荷増加に伴うシステム設計の見直し等の作業が発生する。このため従来の機能の導入のためには所定の期間を要し、費用が発生するという導入障壁があった。

【００１５】また、リモートセンターとの通信回線の容量が十分でない場合に非同期型リモートコピー機能を行うと、リモートセンターへ未反映の更新データが増大するという課題があった。

【００１６】本発明の目的は、新規ソフトウェアの導入を必要とせず、ディスクサブシステムの機能のみで、更新の順序性やデータの整合性を保証でき、導入が容易かつメインセンターの性能低下が少ない、非同期型のリモ

ートコピー機能を実現することである。

【００１７】本発明の別の目的は、非同期型のリモートコピー機能を大容量のデータ格納を行えるディスクサブシステムに適用することで、ディスクサブシステムの顧客に高額の回線使用料を負担させることなく、リモートコピー機能を実現することにある。

【００１８】

【課題を解決するための手段】相互に遠隔地に存在するメインセンターとリモートセンターに、それぞれゲートウェイとなるディスクサブシステム（以下、ゲートウェイサブシステム）を一台ずつ配置し、ゲートウェイサブシステムをデータ伝送線路に接続する。そして、両センターのリモートコピーを実施する必要があるディスクサブシステム全てを、センター内のそれぞれのゲートウェイサブシステムに接続する。

【００１９】メインセンターのディスクサブシステムのリモートコピーの対象となるボリュームと、メインセンター内のゲートウェイサブシステムの任意のボリュームとの間は、同期型リモート機能により接続し、データの二重化を行う。これによりメインセンター内のリモートコピー対象のディスクサブシステムのボリュームと、メインセンター内のゲートウェイサブシステムのボリュームにおいて、システムの処理時間の遅れ等を無視できれば、同一のデータが保持される。

【００２０】メインセンターとリモートセンターのゲートウェイサブシステムの各ボリュームの間は、非同期型のリモートコピーによりデータの二重化を行う。ただし、メインセンターのゲートウェイサブシステムは、自己のサブシステム内のボリュームが更新された順番に従い、更新データをリモートセンターのゲートウェイサブシステムに送付し、リモートセンターのゲートウェイサブシステムは受け取った順番に従い、更新データを自己のサブシステム内のボリュームに反映する。

【００２１】リモートセンター内のゲートウェイサブシステムと各ディスクサブシステムの各ボリュームの間は同期型リモートコピーによりデータの二重化を行う。これにより、リモートセンター側のゲートウェイサブシステムのボリュームとリモートコピー対象のディスクサブシステムのボリュームにおいて、巨視的にみて常に同一のデータが保持される。

【００２２】なお、ゲートウェイサブシステムでは、リモートコピー対象のボリュームのデータは自己のゲートウェイサブシステム内のバッファメモリに格納される。従って、バッファメモリ以外に、データ格納用のサブシステムのエリアは通常は必ずしも必要としない。但し、利用可能なサブシステムのエリアがあれば、伝送線路を経由したデータ送受の際に、伝送線路の容量に応じて必要となるサブシステムのエリアを利用することはできる。

【００２３】以上の構成により、ディスクサブシステム

の機能のみで、メインセンターの複数のディスクサブシステムと、リモートセンターの複数のディスクサブシステムのデータの二重化を更新の順序性を保ちながら実現できる。リモートセンターへの更新データの反映は、メインセンターの各ディスクサブシステムの更新処理とは非同期に実施することができる。これにより高性能で導入の容易な災害バックアップシステムを提供することができる。また、伝送線路の通信容量に応じて、適宜、サブシステムのエリアを用いることができ、顧客の回線使用料負担を軽減できる効果がある。

【００２４】

【発明の実施の形態】以下、図面を参照しながら本発明を汎用コンピュータシステムに適用した場合の一例について説明する。

【００２５】図１に、汎用コンピュータシステムを装備した複数のデータセンターにおいて、任意の２つのセンター間でデータの二重化を行うために、本発明を適用したときの構成例を示す。

【００２６】メインセンター側の一台又は複数台のディスクサブシステムと、リモートセンター側の一台又は複数台のディスクサブシステムは、ホストコンピュータを介さずに、ゲートウェイサブシステムを介して接続され、両センター間でデータの二重化を行うリモートコピーシステムを実現している。

【００２７】図１のメインセンター１２において、中央処理装置（ホストコンピュータ）１は、インタフェースケーブル２を介して、ディスクサブシステム３－１、３－２、……３－ｎに接続されている。ディスクサブシステム３－１、３－２、……３－ｎは、ホストコンピュータ１から参照又は更新されるデータを格納する。ゲートウェイサブシステム５は、インターフェースケーブル４を介して、ディスクサブシステム３－１から３－ｎと接続される。

【００２８】ゲートウェイサブシステム５は、ホストコンピュータがディスクサブシステム３－１等にデータの書込み要求を発行すると、これに同期して当該データを自己のサブシステム内のバッファメモリにも書込む。更に、自己のサブシステム内のバッファメモリにデータが書込まれたこととは非同期に、遠隔地に存在するゲートウェイサブシステム７に対し、データの書込み指示を行う。ゲートウェイサブシステム５は、ディスクサブシステム３－１から３－ｎの台数にかかわらず、必ず一台で構成される。

【００２９】ゲートウェイサブシステム７は、リモートセンター１３に設置され、インタフェースケーブル６を介して、メインセンター１２のゲートウェイサブシステム５と接続されている。なお、インタフェースケーブル６は、一般の通信回線と接続することも可能である。本例ではこの点も含めてインタフェースケーブル６として記述する。ゲートウェイサブシステム７は、ゲートウェ

イサブシステム５から受け取ったデータを、書込み指示のあった順に、自己のサブシステム内のバッファメモリに格納する。ゲートウェイサブシステム７は必ず一台で構成される。

【００３０】ディスクサブシステム９－１、９－２、……９－ｎは、インタフェースケーブル８を介して、ゲートウェイサブシステム７と接続される。ディスクサブシステム９－１等は、メインセンター１２からゲートウェイサブシステム７にデータの書込み要求があった場合には、これに同期して当該データを自己のサブシステム内にも書込む。

【００３１】つまり、ホストコンピュータ１から一台または複数台のディスクサブシステム３－１から３－ｎに対しデータの書込み指示があった場合には、リモートセンター１３内の一台または複数台のディスクサブシステム９－１から９－ｎにも同じデータが格納される。図１の矢印は、ホストコンピュータ１から書込み指示のあったデータの流れを示している。

【００３２】ホストコンピュータ１１は、リモートセンター１３においてディスクサブシステム９－１から９－ｎとインタフェースケーブル１０によって接続され、ディスクサブシステム９－１等に対し、参照及び更新を行う中央処理装置である。ホストコンピュータ１１は、メインセンター１２のホストコンピュータ１が災害や故障等により本来の機能を果たせなくなった場合に、ホストコンピュータ１の代替となって処理を行うことが出来る。このほか、ディスクサブシステム９－１等に格納されているデータを使用して、メインセンター１２のホストコンピュータ１とは異なる処理を、ホストコンピュータ１とは別個独立に実行することができるものである。但し、ホストコンピュータ１１がディスクサブシステム９－１等に対し処理を行わない場合には、ホストコンピュータ１１は不要である。

【００３３】本発明の実施の形態として、データの二重化方法と運用の概略を図２、図３を用いて説明する。

【００３４】二重化の対象となるデータが格納されたボリュームやデータセット、ディスクサブシステムは、事前に運用者が選択する。そして、対象ボリュームや対象データセット及びディスクサブシステムと、選択したデータの複製を格納するボリュームやデータセット及びディスクサブシステムとの関係を、予め運用者がホストコンピュータからディスクサブシステムに対し設定しておく。

【００３５】上記の選択、設定に際し、専用のコンソールやサービスプロセッサを接続又は装備できるディスクサブシステムの場合には、ホストコンピュータを利用せず、そのコンソールやサービスプロセッサを通じて設定できる。図２のフローはホストから選択・設定を行う場合を示している。

【００３６】設定方法としては、上記のボリュームやディスクサブシステムを意味する具体的なアドレスを指定する方法や、ディスクサブシステム内の制御プログラムによって、アドレスの任意の範囲から選択する方法をとることもできる。初期設定として、パス設定やペア設定を行う例を示してある（図２、２０１）。

【００３７】ホストコンピュータ１（図１）から、ディスクサブシステム３－１、３－２、……、３－ｎ（２１１）に対し、データの書込み要求（以下、ライトコマンド）が発行される（図２、２０２）と、ディスクサブシステム３－１、３－２、……、３－ｎはライトコマンドにもとづき自己のディスクサブシステム内へデータ格納処理を実行しつつ、ゲートウェイサブシステム５（２１２）に対し、そのデータのライトコマンドを発行する（２０３）。ここで、ライトコマンドとは、データを書込むための指示と書込みデータそのものとを転送するコマンドである。

【００３８】ゲートウェイサブシステム５は、ライトコマンドを受領するとライトコマンドに対する処理を実行する（２０４）。自己のゲートウェイサブシステム内のバッファメモリへのライトコマンドに対するデータ格納処理が完了すると、ゲートウェイサブシステム５は処理の完了をディスクサブシステム３－１、３－２、……、３－ｎ（２１１）に報告する。これに伴い、処理が完了した順にライトコマンド番号をライトコマンド毎に付与しておき（２０５）、自己のサブシステムの処理能力に基づいて決定された契機で、ライトコマンド番号が付与されたライトコマンドを、ライトコマンド番号順にゲートウェイサブシステム７（２１３）に対し発行する（２０６）。

【００３９】ディスクサブシステム３－１、３－２、……、３－ｎ（２１１）は、ホストコンピュータ１より発行されたライトコマンドに対し、そのライトコマンドに対する処理、即ち、自己のサブシステム内へのデータ格納処理が完了し、かつ、ゲートウェイサブシステム５（２１２）から書込み処理の完了が報告されていること（２２１）を条件に、ホストコンピュータ１に対しライトコマンドに対する処理の完了報告（２２２）を行う。

【００４０】ゲートウェイサブシステム７（２１３）は、ゲートウェイサブシステム５（２１２）から発行されたライトコマンドに付与されているライトコマンド番号により、付与された番号順にライトコマンドを受領していることを確認すると、ライトコマンドに対する処理、即ち、自己のサブシステム内のバッファメモリへのデータ格納処理（３０１）を行う。これに伴い、ディスクサブシステム９（３１１）に対し、そのデータのライトコマンドを発行する（３０２）。ディスクサブシステム９（３１１）は、ゲートウェイサブシステム７から発行されたライトコマンドを受領すると、ライトコマンドに対する処理、即ち、自己のサブシステム内へのデータ格納処理を実行する（３０３）。

【００４１】ディスクサブシステム９－１、９－２、……、９－ｎ（３１１）は、ライトコマンドに対する処理、即ち、自己のサブシステム内のバッファメモリへのデータ格納処理が完了すると、ゲートウェイサブシステム７に対し、処理の完了報告（３２１）を行う。ゲートウェイサブシステム７（２１３）は、自己のサブシステム内へのデータ格納処理が完了し、かつ、ディスクサブシステム９－１、９－２、……、９－ｎから書込み処理の完了が報告されていることを条件に、ゲートウェイサブシステム５に対し、ライトコマンドに対する処理完了報告（３２２）を行う。

【００４２】本発明により、ホストコンピュータ１から書込まれたデータは、ディスクサブシステム３－１、３－２、……、３－ｎと、ゲートウェイサブシステム５の間で二重化され、巨視的にみて常に同一の状態に保たれる。この際にゲートウエイサブシステム５において更新の順序を保持するための情報（通番）が付加される。

【００４３】また、ゲートウェイサブシステム５とゲートウェイサブシステム７との間は更新の順序を保証しながら非同期のリモートコピーでデータの二重化が行なわれる。ディスクサブシステム９－１、９－２、……、９－ｎはゲートウェイサブシステム７の更新に同期してデータが更新される。これらはすべてゲートウエイ機能を有するディスクサブシステムを含めて、ディスクサブシステムの機能のみで実現され、ホストコンピュータの処理能力に対し負担とならない。

【００４４】図４に、各ゲートウェイサブシステム内のバッファー領域を用いて、伝送線路の通信容量が十分でない場合の動作を説明する。同一の符号は既に説明済みであることを示す。このシステムでは、書込み要求のあったデータを一時的に格納するバッファー領域を、各ゲートウエイサブシステム内に設けておく。通常の伝送線路におけるバッファメモリが溢れることを防止するためである。かかるバッファー領域に格納されたサブシステムのデータは、伝送線路を介してメインセンターからリモートセンターへ送られ、そこでリモートセンター側のバッファー領域を介して、ゲートウエイサブシステムへ入力される。こうすれば、二重化の時間的一致度は減少するものの、大容量の通信回線を使用せずとも、非同期型リモートコピー機能を実現できる。

【００４５】

【発明の効果】新規ソフトウェアの導入を必要とせずディスクサブシステムの機能のみで、更新の順序性やデータの整合性を保証でき、導入が容易でかつメインセンターの処理性能の低下が無い非同期型のリモートコピーシステムを実現できる。

【００４６】また、伝送線路の通信容量に応じて、適宜、サブシステムのエリアを用いることができ、顧客の回線使用料負担を軽減できる効果がある。

【図面の簡単な説明】

【図１】本発明の一実施の形態におけるリモートコピーシステムの全体構成を示す図である。

【図２】リモートコピーシステムの処理の詳細なフローチャートを示す図である。

【図３】図２の続きである、リモートコピーシステムの処理の詳細なフローチャートを示す図である。

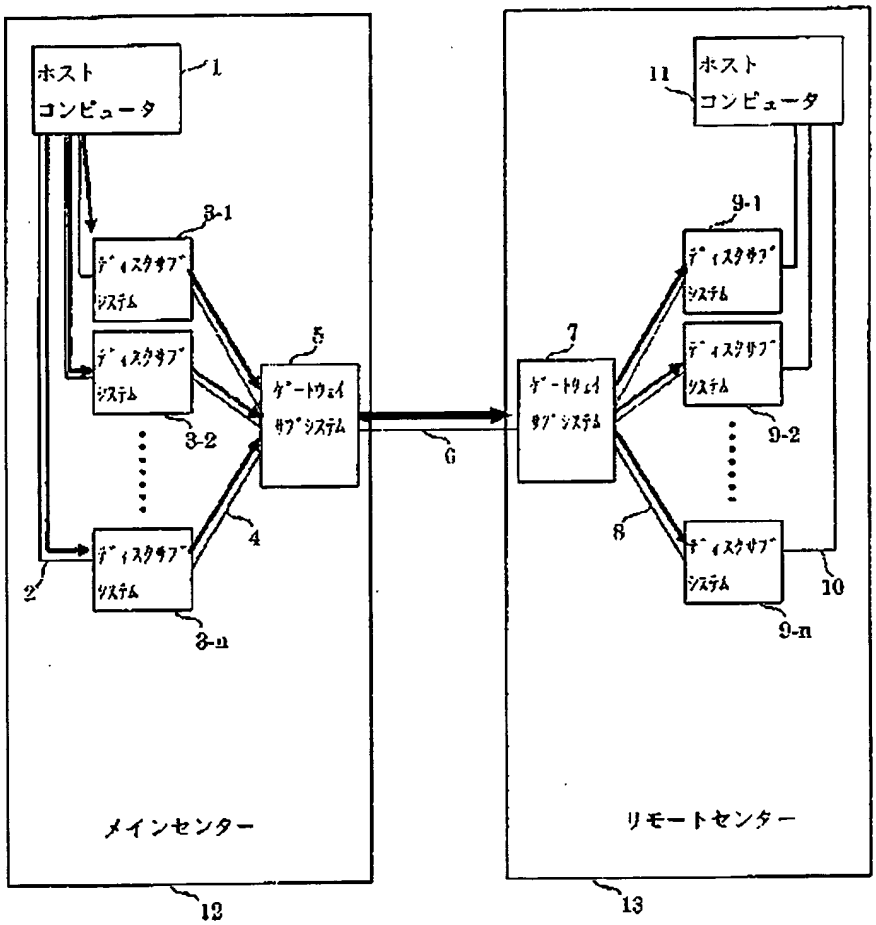
【図４】ゲートウエイサブシステム内にバッファ領域を設けた場合のリモートコピーシステムの処理のフローチャートを示す図である。

【符号の説明】

１…ホストコンピュータ、２…インタフェースケーブル、３…ディスクサブシステム、４…インタフェースケーブル、５…ゲートウェイディスクサブシステム、６…インタフェースケーブル、７…ゲートウェイディスクサブシステム、８…インタフェースケーブル、９…ディスクサブシステム、１０…インタフェースケーブル、１１…ホストコンピュータ、１２…メインセンター、１３…リモートセンター。

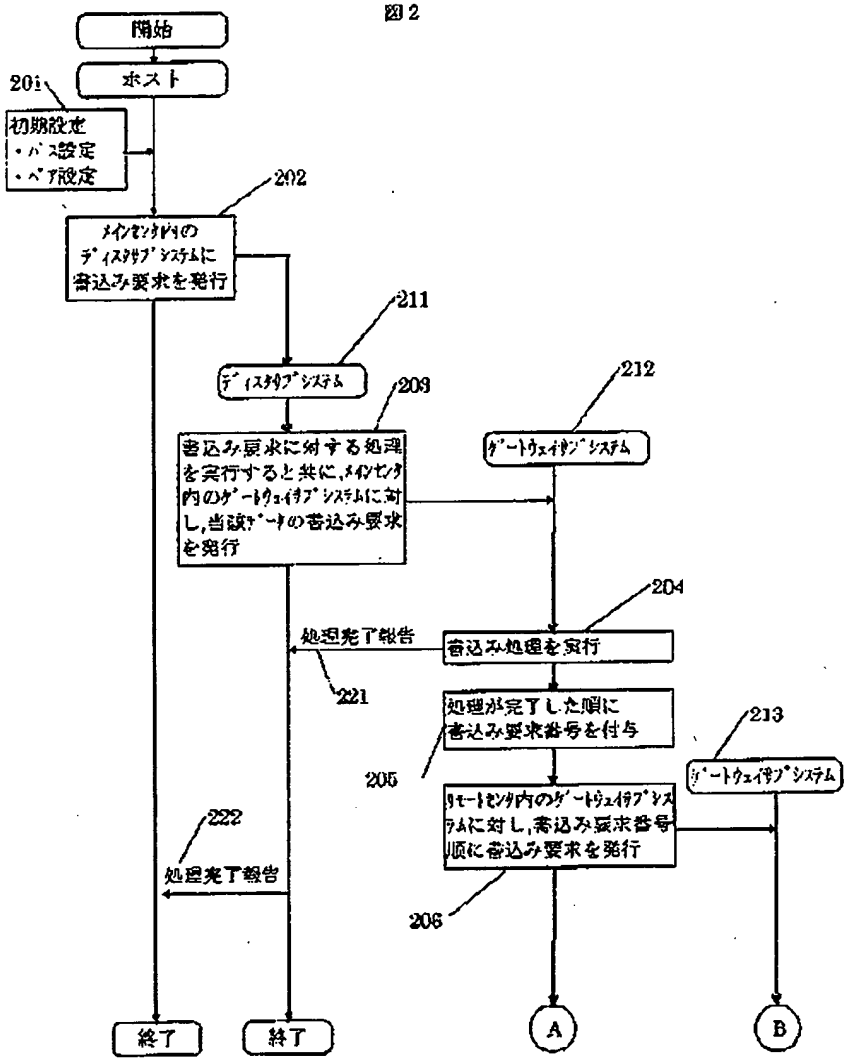
【図1】

図1


ホスト
コンピュータ
1
3-1
3-2
3-n
2
4
5
6
7
8
9-1
9-2
9-n
10
11
ホスト
コンピュータ
メインセンター
リモートセンター
12
13

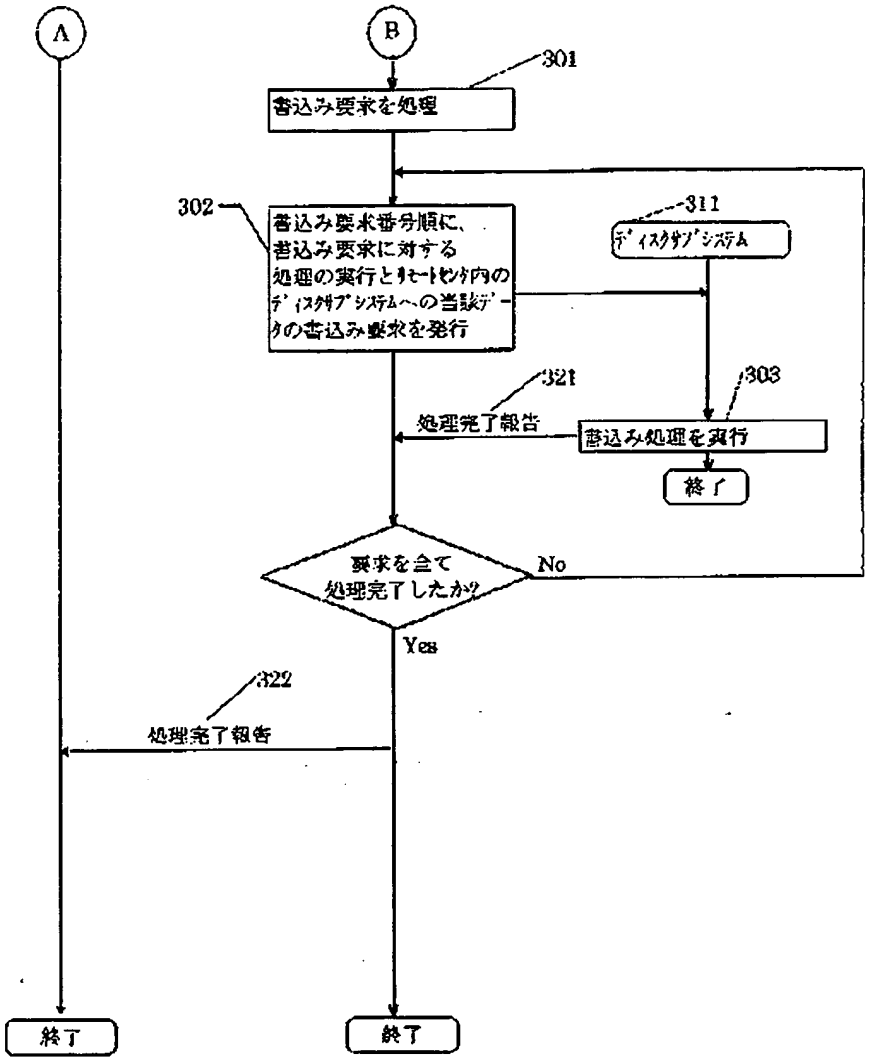
【図２】

図２


開始
ホスト
201
初期設定
・パス設定
・ペア設定
202
メインセンタ内の
ディスタサブシステムに
書込み要求を発行
211
ディスタサブシステム
209
書込み要求に対する処理
を実行すると共に，メインセンタ
内のゲートウェイサブシステムに対
し，当該データの書込み要求
を発行
212
ゲートウェイサブシステム
204
処理完了報告
書込み処理を実行
221
処理が完了した順に
書込み要求番号を付与
213
205
ゲートウェイサブシステム
リモートセンタ内のゲートウェイサブシス
テムに対し，書込み要求番号
順に書込み要求を発行
222
処理完了報告
206
終了
終了
A
B

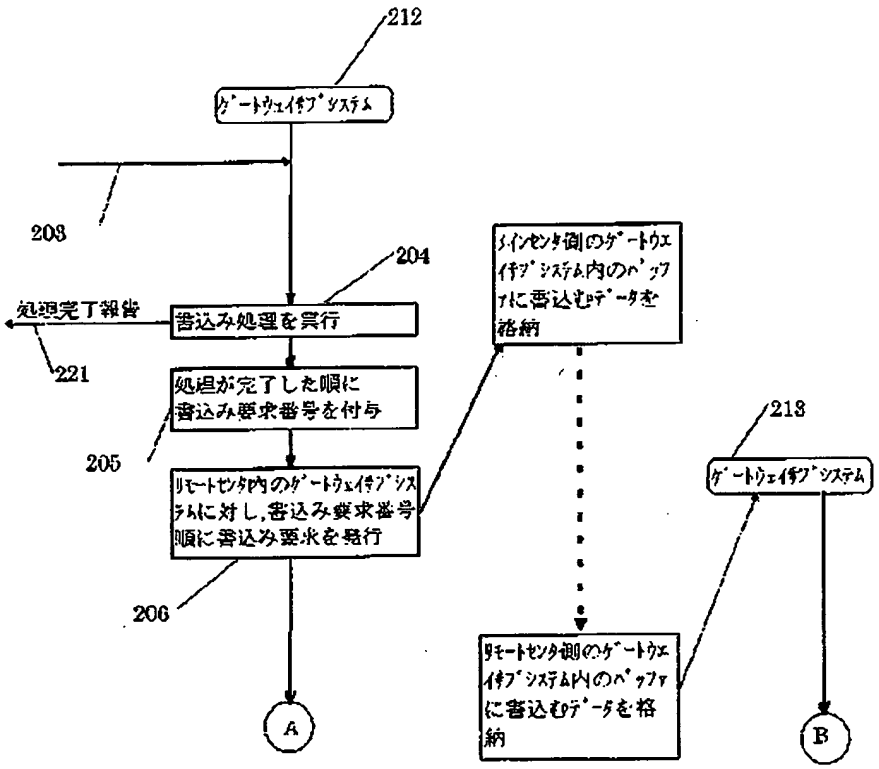
【図３】

図３


A
B
301
書込み要求を処理
302
書込み要求番号順に、書込み要求に対する処理の実行とリモートセンタ内のディスクサブシステムへの当該データの書込み要求を発行
311
ディスクサブシステムA
321
処理完了報告
303
書込み処理を実行
終了
要求を全て処理完了したか？
No
Yes
322
処理完了報告
終了
終了

【図４】

図４



フロントページの続き

(72)発明者 島田 朗伸
神奈川県小田原市国府津2880番地 株式会社日立製作所ストレージシステム事業部内

Ｆターム(参考) 5B018 GA04 HA05 MA12 QA01
5B034 AA01 BB17 CC02 DD06
5B065 BA01 CA50 CC08 CE22 EA23 EA35
5B082 DA02 DE03 GB02 GB06 HA03 HA10
5B083 AA02 AA09 CD11 CE01 DD13 EE08 GG05